图书情报工作
第63卷 第3期 2019年2月

数字图书馆知识发现的数据驱动机制及绩效优化研究*

李洁[1] 毕强[1] 许鹏程[1] 牟冬梅[2]
[1] 吉林大学管理学院 长春 130022 [2] 吉林大学公共卫生学院 长春 130021

摘要：[**目的/意义**]数据驱动环境下,探讨数字图书馆知识发现平台的数据驱动机制和优化方案有利于从方法论认识层面为其供给侧改革提供理论支持。[**方法/过程**]借助系统动力学方法,通过仿真呈现数字图书馆知识发现的数据驱动的动力形成机制;从绩效优化视角,运用粒计算方法为其驱动优化提供可行方案。[**结果/结论**]影响数字图书馆知识发现的数据驱动因素主要包括数据维度、语义关联维度、可视化维度和价值维度,从维度的形成和绩效作用关系看,数字图书馆知识发现的数据驱动是一个螺旋式发展的动态系统,其绩效优化的关键点就在于数据的知识价值开发程度,经实证研究,将知识粒度作为实现其优化的切入点能较好地提升数字图书馆知识发现的数据驱动效果。

关键词：数字图书馆 知识发现 数据驱动
分类号：G250.76
DOI:10.13266/j.issn.0252-3116.2019.03.001

1 引言

21世纪初期,S. C. William [1]将统计学领域的数据研究拓展到先进计算领域,探讨了数据科学的六大类技术范畴,从独立学科视角系统论述了数据科学。随着大数据激发的DT时代的初级转型,基于“数据”解决“问题”的新一代数据驱动方法论认识成为目前数据科学研究范畴的新焦点[2]。数据驱动方法论从问题到数据又回归问题的认识思路体现了第四代研究范式——数据密集型科学发现范式的核心特征:物理对象不再是研究的直面者,研究人员只需面向数据挖掘所需的信息和知识[3]。这一核心特征,使科学研究人员更接近问题形成的本源。从数据直面用户、服务和管理的数据驱动思维为数字图书馆管理理念和服务体系的整体转变提供了契机。直面用户,陈臣等[4]指出图书馆个性化服务应重视数据驱动的数据化作用,为用户提供基于小数据分析的精准画像。直面服务,张晓林[1]针对数据时代的数据化新常态,主张从民生短板和国情痛点出发进行知识服务供给侧的结构性改革。王世伟[5]则基于数据驱动在国家创新战略、规划政策及图情事业中的基础战略地位和助推作用,提出图情教育的创新转型因应数据驱动而为、而谋、而动。王丹丹[6]倡导发挥数据驱动在图书馆建设中的优势,综合利用本地数据和联合数据推动馆藏的系统化组织。承继前人的研究成果,本文着眼于数字图书馆知识发现应用中的数据驱动,应用系统动力学方法分析数字图书馆知识发现的数据驱动动力形成机制,从数据驱动的方法论认识层面为数字图书馆知识服务的供给侧改革提供理论支持;运用粒计算方法,从影响知识价值开发的源头出发,为数字图书馆知识发现的数据驱动提供具体的优化方案,进而通过数字图书馆知识发现数据驱动绩效的价值提升,在深化数字图书馆数据化知识组织的基础上,推进数字图书馆知识服务供给侧在数据化向智能化螺旋进阶中的改革进程。

2 数字图书馆知识发现的数据驱动维度

数字图书馆的知识发现服务程式是在数据碎片化、语义关联化、知识可视化综合驱动下实现系统力化形态由数据输入到知识输出的转变过程,以碎片化为核心的数据维度驱动,以深度语义、广度关联为导向的

* 本文系国家自然科学基金面上项目“嵌入式知识服务驱动下的领域多维知识库构建”(项目编号:71573102)研究成果之一。
作者简介：李洁(ORCID:0000-0002-3929-729X),博士研究生,E-mail:lijie16@mails.jlu.edu.cn;毕强(ORCID:0000-0001-7381-4986),教授,博士生导师;许鹏程(ORCID:0000-0001-5519-8550),硕士研究生;牟冬梅(ORCID:OOOO-0003-0237-034X),教授,博士生导师。
收稿日期:2018-07-08 修回日期:2018-10-10 本文起止页码:6-13 本文责任编辑:易飞

语义关联维度驱动,以视觉表征、意义构建和推理预测为主要内容的可视化维度驱动以及以数据柔术性和数据洞见性为知识转化衡量标准的价值维度驱动构成了数字图书馆知识发现的驱动逻辑主线,具体的驱动维度以及所涉及的要素和属性如下:

2.1 数据维度

数字图书馆的知识发现是从数据直面问题,实现不同结构(结构化、半结构化和非结构化)数据的碎片化、语义关联化、可视化、知识化的递阶解构,致力于数据向知识资本转化的挖掘、解析、创造与领域知识应用。在结构化导向的数据维度层面,主要涉及以下要素:①数据质量,数字图书馆知识发现的数据维度质量属性包括数据源的正确性和完整性以及数据的一致性、连续性、时效性、精确性、自描述性和形式化程度,还有数据的精度和同步程度等;②数据结构,是指可以用二维表结构来逻辑表达实现的程度,数字图书馆文献知识发现的对象是非结构化数据,数据库知识发现的对象是结构化数据;③数据相关性,指的是数据的相关性分析方式,数据驱动下的数字图书馆知识发现主要通过数据的相关性分析而非因果分析去分析和解决问题,数据本身的相关性程度与特征对隐性知识的挖掘和规律发现具有重要作用;④数据加工程度,是指数据的预处理程度,一般涉及数据的审计、清洗、变换、抽象、集成、标注、排序等要素。

2.2 语义关联维度

数据驱动下数字图书馆的知识发现服务是面向具有语义分析和关联特性的"语义互联网"的,以语义概念、语义类型、语义关系、语义标注、语义推理为核心要素的语义化和以关联结构、关联强度和关联规则为核心要素的关联化融合驱动着数字图书馆知识发现的实现,两者交叉作用下的语义关联驱动维度主要涉及以下内容:①语义概念,其是描述数据本身的具有语义属性的概念单元,相同属性的语义概念通过聚合生成语义类型;②语义关系,对数据之间的内在联系的呈现,通过逻辑三元体的语义角色和动词核心作用进行呈现,LSR(labeled semantic relations)作为带标记的语义关系既能呈现概念间关系,又能呈现关系类型;③语义标注,通过语义元数据将实体的概念、属性和关系等值与相应的语义描述进行关联语义化的过程[7]。④语义推理,是借助特定的意义公设手段,应用语义系统框架揭示词项语义结构和语义关系的推理方法;⑤关联结构,通过数据结构表示或特定方法呈现数据实体的内外部关联特性;⑥关联关系,实体与实体间直接或间接、隐性或显性的结构化关系;⑦关联规则,可以用 X→Y 的蕴涵式表达, X 和 Y 分别是关联规则的先导和后继,具有支持度、信任度和强弱之分,强关联规则是指同时满足用户定义的最小支持度阈值和最小置信度阈值的关联规则。综合来看,有关语义关联维度驱动的要素属性涉及语义相似度、语义距离、语义相关度、语义重合度、语义匹配度、语义粒度、语义隶属度、语义半径、语义权重、语义贴近度、标签广度、链接广度、链接深度、关联维度、关联强度、关联阶度和关联粒度等[8]。

2.3 可视化维度

可视化维度驱动的知识发现既是数据知识网络形态呈现的过程,又是知识分析与预测的知识规律发现和领域知识探索的过程,借助发现平台内嵌的可视化技术以静动态的知识网络图谱为用户呈现满足其目标检索、解读、预测等任务基础上的具有视觉表征和智能化雏形特征的知识域,其主要涉及以下要素:①视觉表征,是指可视化的视觉呈现形式,一般分为呈现为视觉物质材料的表层形式(如形状、色彩、机理等)和呈现空间关系的深层形式(如和谐、对称、均衡、节奏等);②意义构建,是指主体对视觉感知的信息,基于过往经验修正和知识结构演化而形成的新的理解的视觉思维;③知识网络结构,反映著者、机构、期刊等之间的合作关系网络情况以及引文分析的耦合、共被引、共现关系网络结构情况等;④知识网络测度,是指衡量知识网络结构的指标,包括中心势、中心度(具体包括节点中心度、群体中心势、紧密中心度、间距中心度)、平均路径长度、凝聚子群等;⑤时间序列,其是通过时间先后顺序排列反映同一统计指标数值变化情况以呈现实体发展脉络并辅助预测的数值,通过时间序列对数据基于用户空间域进行知识挖掘、聚类和分析更有利于开发平台数据的知识价值。此维度涉及的主要属性包括视觉隐喻度、视觉通道畅通度、视觉突出性、符号化程度、k-核子网、聚类系数、载荷度、引文长度、引文频次、耦合度、主体相关度、主体可控度和主体知识结构等。

2.4 价值维度

数字图书馆知识发现的数据驱动功能在于实现"数据→用户→知识发现"服务空间的良性循环,数据的价值维度作为循环过程中直接接近数据驱动绩效的动力源,关系着数据输入端到知识输出端的转化能力,主要包含以下因素:①数据柔术,指数据转化成数据产品的能力,即成品性、商品性;②数据洞见,即数据能被

发现且自身带有信息价值的程度，其与主体的数据意识、经验和分析能力息息相关；③数据资产，其是数据从资源向资产的转变，可以对数据进行定价、产权归属、交易等[9-10]。

3 数字图书馆知识发现的数据驱动机制

数字图书馆知识发现平台的数据驱动强调的是基于数据解决问题的求解过程中用户需求与服务的解耦，其动力来源是数据维度、语义关联维度、可视化维度和价值维度融合驱动下的各要素，通过对各维度要素动力类型的分析，能够清晰明辨数字图书馆知识发现数据驱动的动力形成机制。

从数字图书馆知识发现数据驱动的各维度要素看，驱动知识发现过程的动力源主要有两种：一种是与数据、语义关联、可视化本身相关的内部动力，即内生动力；一种是受外部条件影响而使各维度要素产生动能的外部动力，即外源动力；基于此，本文从系统动力学的常用动力类型——内生动力和外源动力对数字图书馆知识发现的数据驱动维度要素进行类型划分与归并，具体如表1所示：

表1 数字图书馆知识发现的数据驱动要素和动力类型

维度	要素	动力类型
数据维度	数据质量	内生动力
	数据结构	内生动力
	数据相关性	内生/外源·动力
	数据加工程度	外源动力
语义关联维度	语义概念	内生动力
	语义关系	内生动力
	语义标注	外源动力
	语义推理	外源动力
	关联结构	内生动力
	关联关系	内生动力
	关联规则	外源动力
可视化维度	视觉表征	内生动力
	意义构建	外源动力
	知识网络结构	内生动力
	知识网络测度	内生/外源·动力
	时间序列	外源动力
价值维度	数据柔术	外源动力
	数据洞见	外源动力
	数据资产	外源动力

注：标注·的为侧重的动力类型

在语义关联维度层面，语义概念、语义关系、关联结构、关联关系都是内生动力；语义标注、语义推理、关联规则都是外源驱动力。可视化维度层面，视觉表征和知识网络结构是内生动力；意义构建和时间序列是外源动力；知识网络测度既以内生动力形式驱动于知识发现，也以外源动力形式间接作用于知识发现。价值维度层面，受数据主体的影响，数据柔术、数据洞见、数据资产都以外源作用力形式对数字图书馆的知识发现进行驱动。整体上，数字图书馆知识发现的内生驱动力来源于数据维度和与语义、关联、可视化因素作用的本身，外源驱动力是有关数据化、语义－关联－可视化相关的技术因素本身，此外，与绩效挂钩的数据的价值要素也在数据主体的主观因素变动作用下对知识发现产生间接驱动作用。各维度要素具体的动力驱动双向因果关系见图1。

从力的作用形式来看，内生动力是与各驱动维度要素本身密切相关的动力源，是各要素本身通过内生力量进行直接驱动的作用形式；外源动力是各驱动维度要素以外生力的作用形式产生驱动作用，不直接作用于实体。如表1所示，数字图书馆知识发现各维度要素的数据驱动作用形式较为显著，而各维度在不同要素的影响下其驱动形式呈现非单一状态，除价值驱动维度外，其他驱动维度既是内生力又是外源力。在数据驱动维度层面，数据质量和数据结构都以内生力的形式对数字图书馆知识发现产生直接的数据驱动作用；数据相关性从分析思维角度以外源力形式产生驱动作用，从相关程度角度以内生力形式直接驱动于知识发现；数据加工程度以外源作用形式间接作用于知识发现。

如图1所示，数字图书馆知识发现的数据驱动是内生动力和外源动力综合作用的结果。数据维度驱动的数字图书馆知识发现是以内生动力为主，且除数据质量和数据结构外，各驱动要素间的关系呈正向性，总体的驱动作用力大小受到数据质量、数据结构、数据加工程度和数据相关性的影响，并且受数据加工程度的影响较大。语义关联维度驱动的数字图书馆知识发现受内生力和外源力的双重影响，且语义概念、语义关系、关联关系和关联规则的驱动作用力更大，除影响语义关系的可用性和一致性属性外，其他因素间及属性间的作用关系均成正向性。可视化维度和价值维度分别驱动的数字图书馆知识发现则更多地提供外源动力，其作用力大小受知识网络结构、知识网络测度、意义构建和数据资产的影响较大。

就各维度作用的数字图书馆知识发现数据驱动形成动力关系而言，系统动力学的结构化思维贯穿整个驱动系统空间。其中，数据维度是内动力和外源动力

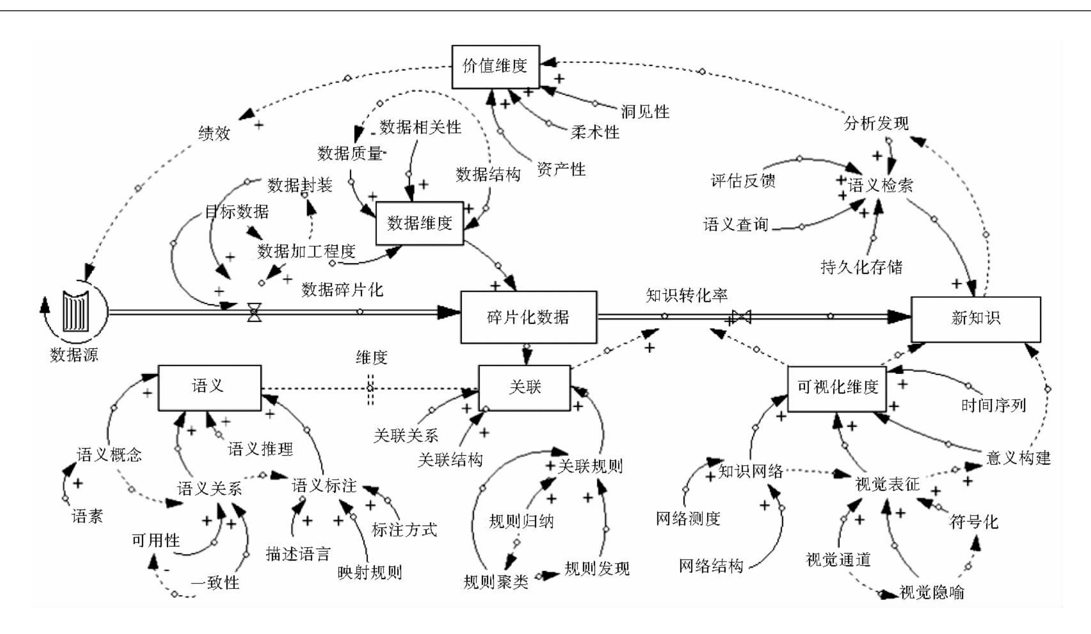


图 1　数字图书馆知识发现数据驱动形成动力因果关系

发生作用的基础,关系着数据碎片化的实现;语义关联和可视化维度是内生动力和外源动力双重驱动的技术中间件,共同作用着知识的转化率;价值维度是内生动力和外源动力发生作用的直接绩效体现,通过反馈将驱动的出发点和落脚点相关联。

4　数字图书馆知识发现数据驱动的绩效优化

4.1　数字图书馆知识发现数据驱动的绩效优化路径

数据科学的基本任务之一就是探索一种能够用于描述数据结构性质的方法工具,结合信息论中对具有知识初步性质的信息的界定——比特加语义的组合(即给一连串比特构成的消息附上语义)、知识表示转换和知识结构转换的等价论说以及数字图书馆知识发现数据驱动场域空间的结构性特征,不难发现:数据和知识的结构性能开发是提升数据-知识转化率的关键。所以,数字图书馆知识发现数据驱动系统空间的绩效优化应从结构主义视角进行驱动优化。而在知识的结构化表达中,知识的结构化程度是由知识粒度开始的,粒结构关系着知识的结构,其能基于知识网络的耦合性质将复杂的知识网络转换为一个较为简单的多层次结构。借助粒计算的多粒度层级方法,本文将通过数据的粒化实现粒化数据向粒化知识的转化,通过知识结构的优化开发提升数字图书馆知识发现的效能。具体的绩效优化思路如图 2 所示:

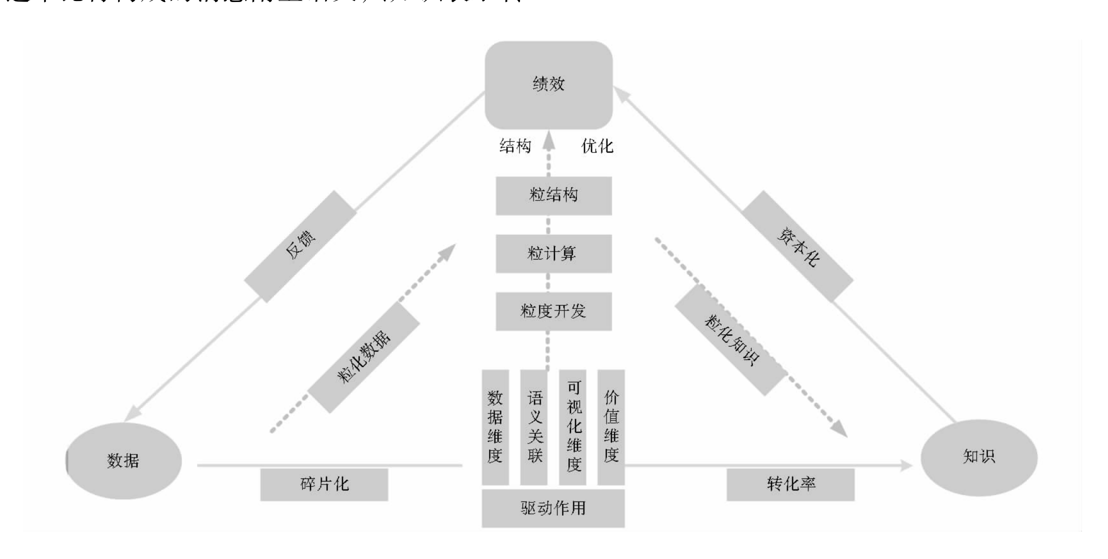


图 2　数字图书馆知识发现数据驱动的绩效优化路径

4.2 数字图书馆知识发现数据驱动的绩效优化方法

粒计算能够基于粗糙集、模糊集、商空间和云模型的融合,以多层次、多粒度思维实现复杂问题的简单求解。针对数字图书馆知识发现数据驱动系统空间的结构化制约特性,采用基于粒结构的粒计算有利于顺应数据科学“基于数据解决问题”的思潮,通过知识粒化的结构性能提升加快数据向知识的转化效率,进而推进数字图书馆知识发现数据驱动的绩效优化。采用粒计算进行问题求解时,多粒度分析是较为有效的方法之一。多粒度分析能够赋予数据以语义粒度属性,推进粒化数据向粒化知识的语义进阶,加快以知识粒度为单元的粒空间构建,进而实现知识的深度挖掘、发现与利用[11-12]。现阶段,我国主流的数字图书馆知识发现平台主要是面向科研用户并为其提供语义检索服务,结合这一现状,本文以知识发现平台的语义检索决策为研究对象,进行具体的数据驱动绩效优化实现,将优化的焦点聚焦于发现平台中用户语义检索辅助决策的功效上,结合多粒度粗糙集理论,采用多粒度分析进行检索决策的服务绩效优化。在应用多粒度分析方法时,按照求同存异的原则,从乐观多粒度融合视角进行多粒度决策的绩效判别和优化[11]。

在多粒度粗糙集的数据建模中,信息系统的形式描述与关系数据库相似[13]。设信息系统 $S=(U,AT,f)$,$f_a: U\rightarrow V_a$ 表示 U 与 V_a 间的关系,即对于任意的 $a\in AT$,V_a 是属性 a 的值域。对于任意的 $x\in U$,x 的信息向量表示为[14]:$f_{(x)}=\{(a),f_a^{(a)} \mid a\in AT\}$。在数字图书馆知识发现数据驱动场域中,平台所提供的语义检索渠道是一个完备的知识发现决策辅助系统,根据以上的公式定义,设数字图书馆知识发现数据驱动场域中用户语义检索决策系统为:

$$DS=\{IS_i \mid IS_i=(U,AT_i,\{V_a\}_{a\in ATi,fi})\},$$

$X_{il}\in U/AT_j$, $i=1,2\cdots,m$, $l=1,2,\cdots,t_i$, $Y_j\in U/\{d\}$, $j=1,2,\cdots,k$

依据乐观多粒度粗糙集求同存异(即非排他性)的核心思想,运用 $des(X_{il})$ 和 $dos(Y_j)$ 对 DS 进行 X_{il} 和 Y_j 的等价描述,则 Z_{ij}^o 表示的多粒度融合规则可以定义为:

$$Z_{ij}^o: \vee_{i=1}^m des(X_{il_i})\rightarrow des(Y_i),\ j=1,2,\cdots,k$$

该规则的确信度 Cer 和支持度 $Supp$ 表示为:

$$Cer(Z_{ij}^o)=\max_{i=1}^m\left\{\frac{|X_{il}\cap Y_i|}{|X_{il}|}\right\},\ Supp(Z_{ij}^o \max_{i=1}^m\left\{\frac{|X_{il}\cap Y_i|}{|U|}\right\};$$

根据多粒度粗糙集的性质规则,数字图书馆知识发现数据驱动场域中用户语义检索决策的乐观多粒度粗糙集确信度的最大值和最小值分别为 1 和 0。

基于对一组局部变量因不一致性而难以有效进行结果比较的情况,应用整体性思维进行判别是常用的方法,基于此,本文从全局决策的角度,通过整体确信度和整体支持度对系统的绩效进行判别[16]。用 $\partial^\circ(DS)$ 和 $\beta^\circ(DS)$ 分别表示其整体确信度和整体支持度,描述如下:

$$\partial^\circ(DS)=\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{1}{k}\sum_{j=1}^{k}Cer(Z_{ij}^o)\times Supp(Z_{ij}^o),$$

$$\beta^\circ(DS)=1-\frac{4}{|U|}\sum_{i=1}^{|U|}\frac{1}{k}\sum_{j=1}^{k}Cer(Z_{ij}^o)\times(1-Cer(Z_{ij}^o))$$

4.3 数字图书馆知识发现数据驱动的绩效优化实现

以数字图书馆知识发现数据驱动的检索结果绩效优化为目标,本文随机选取 7 名吉林大学“鼎新中文发现”的学生用户作为实验受试对象,进行基于“粒结构”理论优化思想的用户检索决策绩效测验,用户专业背景不限,检索内容不限,文献类型选择期刊,年份限定近五年。数字图书馆知识发现数据驱动场域为用户进行知识检索提供了多种渠道,以吉林大学数字图书馆知识发现平台为例,高级检索窗口提供了全部字段、主题、摘要、标题、关键词等不同粒度等级的检索渠道;结果输出方面,提供了出版日期、馆藏、学术性、相关性和引文量等排序方式;在可视化呈现方面,以相关性为主要呈现依据,提供了知识点、作者和机构等知识图谱类型。本文以关键词、标题和摘要 3 种不同粒度等级的检索渠道为条件属性源,每个条件源下将系统所提供的相关性、引文量和学术性作为考核因素,将用户对检索结果的满意和不满意态度作为绩效属性进行“求同存异”的多粒度检索决策的绩效测验。具体表述如下:

基于多粒度结构优化的数字图书馆知识发现检索平台是一个完备的决策系统,定义为 $DS=\{IS_i \mid IS_i=(U,AT_i,\{V_a\}_{a\in ATi,fi},D)\}$,受试者的 3 次检索结果记录表示为 $U=(e_1,e_2,e_3,e_4,e_5,e_6,e_7)$,条件属性 D_1、D_2、D_3 分别为知识发现平台关键词、标题、摘要检索渠道源的具体考核因素——相关性、引文量与学术性,绩效属性为 $D=\{Y,N\}$。Y 为满意,N 为不满意,得到的具体粒结构空间见表 2。

基于粒空间可得:

$$G_1=\{\{e_1,e_3,e_4,e_5,e_7\},\ \{e_2,e_6\}\},$$

表 2　多渠道检索决策粒结构空间集

源 / 对象	D1			D2			D3			绩效
属性	相关性	引文量	学术性	相关性	引文量	学术性	相关性	引文量	学术性	D
e 1	低	高	高	高	低	高	低	低	高	N
e 2	高	高	高	高	低	高	高	低	高	Y
e 3	低	高	高	低	高	高	高	高	高	Y
e 4	低	高	高	低	高	高	低	高	高	N
e 5	低	高	高	高	高	高	高	高	高	Y
e 6	高	高	高	高	低	高	高	高	高	Y
e 7	低	高	高	高	高	高	低	低	高	N

$G_2 = \{\{e_1, e_2, e_6), \{e_3, e_4\}, \{e_5, e_7\}\}$,

$G_3 = \{\{e_3, e_5, e_6\}, \{e_1, e_7\}\{e_2\}, \{e_4\}\}$

根据决策的绩效属性，得到决策绩效类为：$U\{d\}$ $=\{\{e_2, e_3, e_5, e_6\}\{e_1, e_4, e_7\}\}$,

记各个源 AT_i，$i = 1,2,3$ 下的条件粒分别为：

$X_{11} = \{e_1, e_3, e_4, e_5, e_7\}$, $X_{12} = \{e_2, e_6\}$,

$X_{21} = \{e_1, e_2, e_6\}$, $X_{22} = \{e_3, e_4\}$, $X_{23} = \{e_5, e_7\}$,

$X_{31} = \{e_1, e_7\}$, $X_{32} = \{e_2\}$, $X_{33} = \{e_3, e_5, e_6\}$, $X_{34} = \{e_4\}$

则检索决策的绩效粒为：$Y_1 = \{e_2, e_3, e_5, e_6\}$, $Y_2 = \{e_1, e_4, e_7\}$,

按照 $des(X_{il} \to des(Y_j)$ 的等价描述，3 个源的 RULL 集如表 3 所示：

表 3　多渠道检索决策粒结构空间 RULL 集

第一个源			第二个源			第三个源		
des	Supp	Cer	des	Supp	Cer	des	Supp	Cer
$X_{11} \to Y_1$	2/7	2/5	$X_{21} \to Y_1$	2/7	2/3	$X_{31} \to Y_1$	0	0
$X_{12} \to Y_1$	2/7	1	$X_{22} \to Y_1$	1/7	1/2	$X_{32} \to Y_1$	1/7	1
$X_{11} \to Y_2$	3/7	3/5	$X_{23} \to Y_1$	1/7	1/2	$X_{33} \to Y_1$	1/7	1
$X_{12} \to Y_2$	0	0	$X_{21} \to Y_2$	1/7	1/3	$X_{34} \to Y_1$	0	0
			$X_{22} \to Y_2$	1/7	1/2	$X_{32} \to Y_2$	2/7	1
			$X_{23} \to Y_2$	1/7	1/2	$X_{32} \to Y_2$	0	0
						$X_{33} \to Y_2$	0	0
						$X_{34} \to Y_2$	1/7	1

多源乐观决策绩效规则集为：

$Z^o_{11}: V^3_{i=1}[e_1]AT_i \to Y_1$, $Suup = \frac{1}{7}$, $Cer = 1$,

$Z^o_{21}: V^3_{i=1}[e_2]AT_i \to Y_1$, $Suup = \frac{1}{7}$, $Cer = 1$,

$Z^o_{31}: V^3_{i=1}[e_3]AT_i \to Y_1$, $Suup = \frac{1}{7}$, $Cer = 1$,

$Z^o_{41}: V^3_{i=1}[e_3]AT_i \to Y_1$, $Suup = \frac{2}{7}$, $Cer = \frac{2}{3}$,

$Z^o_{51}: V^3_{i=1}[e_3]AT_i \to Y_1$, $Suup = \frac{2}{7}$, $Cer = 1$,

$Z^o_{61}: V^3_{i=1}[e_3]AT_i \to Y_1$, $Suup = \frac{1}{7}$, $Cer = 1$,

$Z^o_{71}: V^3_{i=1}[e_3]AT_i \to Y_1$, $Suup = \frac{2}{7}$, $Cer = \frac{2}{3}$,

$Z^o_{12}: V^3_{i=1}[e_1]AT_i \to Y_1$, $Suup = \frac{1}{7}$, $Cer = 1$,

$Z^o_{21}: V^3_{i=1}[e_2]AT_i \to Y_2$, $Suup = \frac{1}{7}$, $Cer = 1$,

$Z^o_{32}: V^3_{i=1}[e_3]AT_i \to Y_2$, $Suup = \frac{1}{7}$, $Cer = \frac{1}{2}$,

$Z^o_{42}: V^3_{i=1}[e_3]AT_i \to Y_2$, $Suup = \frac{1}{7}$, $Cer = \frac{1}{3}$,

$Z^o_{52}: V^3_{i=1}[e_3]AT_i \to Y_2$, $Suup = \frac{1}{7}$, $Cer = 1$,

$Z^o_{62}: V^3_{i=1}[e_3]AT_i \to Y_2$, $Suup = \frac{1}{7}$, $Cer = \frac{1}{2}$,

$Z^o_{72}: V^3_{i=1}[e_3]AT_i \to Y_2$, $Suup = \frac{1}{7}$, $Cer = \frac{1}{3}$,

整体确信度为：

$$\partial^\circ(DS) = \frac{1}{2 \times 7}\left(\left(\frac{1}{7} \times 1\right) + \frac{1}{7} \times 1\right) + \frac{1}{7} \times 1) + \frac{2}{7} \times \frac{2}{3}) + \frac{2}{7} \times 1) + \frac{2}{7} \times \frac{2}{3})) = 0.0884$$

整体支持度为：

$$\beta^{\circ}(DS) = 1 - \frac{4}{2\times7}\left(\left(\left(\frac{1}{2}\times\frac{1}{2}\right)+\frac{2}{3}\times\frac{1}{3}\right)+\frac{2}{3}\times\frac{1}{2}\right)=\frac{1}{2}\times\frac{1}{2}\Big)+\frac{2}{3}\times\frac{1}{3}\Big)\Big)=0.73016$$

从整体看来，乐观多粒度决策的整体确信度和支持度都较好。采用十字交叉验证方法，通过聚类系数的个数增减效果比较其决策的合理度，通过聚类分析，得到的数字图书馆知识发现数据驱动场域中用户语义检索的多粒度粗糙集决策绩效合理度如图3所示：

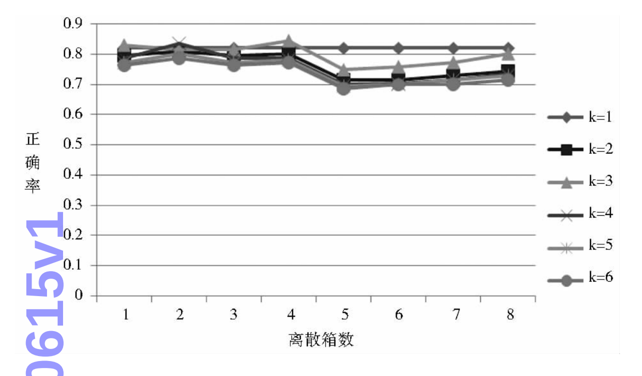


图3 基于多粒度粗糙集方法的数字图书馆知识发现数据驱动的语义检索决策绩效优化效果

由图3所示，分别考虑聚类数从1到6进行用户语义检索途径的正确性预测实验测试，从上文分析结果来看，当K=2,3时，通过乐观决策提升其驱动绩效是较为合理的，多粒度粗糙集在绩效优化中的优势明显，多粒度决策对用户语义检索的知识发现服务辅助效果良好。当数量离散箱数为4时，聚类个数为3时，产生的绩效效果是最佳的。针对7个用户的检索需求，知识发现平台为用户提供综合关键词、标题、摘要三者的检索决策渠道时，用户检索到的知识信息最佳，平台的检索绩效最优。

通过对检索决策的聚类对比，采用单粒度方法是无法较好满足用户语义检索要求的，借助多粒度方法，能更全面、更准确地为用户利用知识发现平台进行语义检索提供决策支持。在多粒度乐观决策下，系统对用户的决策匹配按照“求同存异”能够针对用户不同的检索目标进行多渠道检索决策的匹配与交叉融合，通过关键词、标题、摘要的综合交叉检索，系统能够为用户提供更优化的检索决策支持，帮助用户尽可能搜索到更全面、更精准、更符合其检索需求的知识信息。

本文对数字图书馆知识发现平台检索决策的绩效通过正确性进行绩效效果判定，依据阈值考察粒计算中多粒度粗糙集对数字图书馆知识发现数据驱动绩效优化的实现程度。总体上，应用多粒度粗糙集对数字图书馆知识发现数据驱动进行绩效优化是有效的，采用多粒度进行语义检索对数据向知识转化的速率提升有较好的效果。在用户的语义检索过程中，应用多渠道的综合检索途径能够更好地发挥数字图书馆知识发现平台对数据资源的知识挖掘、整合和分析性能，针对用户的检索需求提供更好的检索结果，提升线上馆藏数据资源向满足用户需求的知识资源的转化效率，提升数据驱动下数字图书馆知识发现的服务能力，进而推进密集型数据科学范式下，由以往受用户碎片化和知识化需求侧牵引的服务推荐向主动式多粒度层级空间的数字图书馆知识服务供给侧的发展转变。

5 结语

大数据环境下，对数字密集型科学的关注正在成为一种趋势，面向学科领域，数字图书馆服务能否从以往粗放型转变至精准型，开启数字图书馆服务转型发展的新方向，实现从资源发现到知识发现的转变，是值得我们认真思考并加以解决的关键问题[15]。数据驱动下数字图书馆知识发现服务的系统力化作用，由输入端数据通过数据化、碎片化、语义化、关联化、可视化的联合驱动，向具有智能化性质的知识进阶。从影响数字图书馆知识发现数据驱动的数据维度、语义关联维度、可视化维度和价值维度的成因关系看，数字图书馆知识发现的数据驱动是一个螺旋式发展的动力生态系统，其优化的最终目标是实现数据资源的知识价值开发。面向价值共创，数字图书馆知识发现的数据驱动优化在综合考虑各因素绩效作用关系的基础上，借助知识结构优化的多粒度方法，能够在数据化向知识化的数字图书馆知识发现服务供给侧结构改革中加速知识价值的转化效率，为用户的语义检索提供更智能的检索决策服务。

参考文献：

[1] WILLIAM S C. Data science: an action plan for expanding the technical areas of statistics[J]. International statistical review, 2010, 69(1):21-26.

[2] 李洁，毕强，张晗，等. 数据驱动下数字图书馆知识发现的服务研究[J]. 情报资料工作，2018(4):6-14.

[3] TOLLE M, TANSLEY D, STEWART W, et al. The fourth paradigm: data-intensive scientific discovery[J]. Proceedings of the IEEE, 2011, 99(8): 1334-1337.

[4] 陈臣，马晓亭. 基于小数据的图书馆用户精准画像研究[J]. 情报资料工作，2018(5):57-61.

[5] 王世伟. 数据驱动的时代特征与图情教育的创新转型[J]. 图

书情报知识,2016(1):15－20.
[6] 王丹丹. 数据驱动的馆藏建设趋势及实践案例[J]. 情报资料工作,2016(5):74－78.
[7] POPOV B, KIRYAKOV A, KIRILOV A, et al. KIM-semantic annotation platform[M]. Berlin: Springer, 2003: 834－849.
[8] 闫晶,毕强,李洁. 数字图书馆资源聚合质量评价指标构建[J]. 图书情报工作,2017,61(24):5－12.
[9] 朝乐门. 数据科学[M]. 北京:清华大学出版社,2016.
[10] 刘士军,鹿旭东,崔立真. 数据化企业制胜之道——数据驱动的创新[M]. 北京:电子工业出版社,2017.
[11] 王国胤,张清华,胡军. 粒计算研究综述[J]. 智能系统学报,2007(6):8－26.
[12] LIN T. Data mining and machine oriented modeling: a granular computing approach[J]. Applied intelligence,2000,13 (2):113－124.
[13] 王国胤,张清华. 不同知识粒度下粗糙集的不确定性研究[J]. 计算机学报,2008(9):1588－1598.
[14] 林国平. 多粒度粗糙计算理论与方法研究[D]. 太原:山西大学,2016.
[15] 毕强,闫晶,李洁. 大数据时代数字图书馆服务转型面临的新形势与新要求[J]. 情报理论与实践,2017,40(12):12－16,5.

作者贡献说明:

李洁:提出研究命题、研究思路,撰写全文;
毕强:提出研究思路;
许鹏程:修改论文;
牟冬梅:统领纲要。

Research on Data Driven Mechanism and Performance Optimization of Knowledge Discovery in Digital Library

Li Jie[1] Bi Qiang[1] Xu Pengcheng[1] Mou Dongmei[2]
[1] School of Management, Jilin University, Changchun 130022
[2] School of Public Health, Jilin University, Changchun 130021

Abstract: [**Purpose/significance**] Under the data-driven environment, exploring the data-driven mechanism and optimization scheme of knowledge discover platform of digital library is conducive to provide theoretical support for supply-side reform from the perspective of methodology. [**Method/process**] By means of the system dynamics method, the data-driven dynamic formation mechanism of digital library knowledge discovery is presented through simulation. From the perspective of performance optimization, the granular computing method is used to provide a feasible solution for its drive optimization. [**Result/conclusion**] The data driving factors that influence the knowledge discovery of digital library mainly include data dimension, semantic association dimension, visualization dimension and value dimension. From the perspective of the formation of dimensions and the role of performance, the data drive of digital library knowledge discovery is a dynamic system of spiral development, the key point of performance optimization lies in the exploitation degree of knowledge value of data. The knowledge granularity as the starting point to achieve its optimization can better improve the data-driven effect of digital library knowledge discovery, according to the experimental studies.

Keywords: digital library knowledge discovery data driven